Actuarial

Data Analytics / Predictive Modeling Seminar Part 3

Trainers:
Paul Bailey
Shawn Balthazar
Jayde Hagen
Kate Oliver
Ben Williams

Please log on to the Azure Remote App and open Emblem

Login: TrainXX@wtwsaas.com

Password: T0wersW@tson16

where XX is the number you were previously assigned; if you don't have a number ask us

ILTCI

Disclaimer

The data used in this seminar is fictitious, and it or results of analyses based on it should not be relied upon for any purpose

Agenda of the Seminar

1. Background to the Workshop
2. Predictive Modeling and Applications to Long Term Care Insurance
3. **Predictive Modeling of Long Term Care Incidence using Generalized Linear Models in Emblem**

Introduction

Where have we come from and why are we here?

- During the conference we discussed how predictive modeling offers alternatives to traditional techniques for developing assumptions
- These alternative techniques can give more accurate predictions, and give better insight into the process being modeled
- On Sunday we ensured that you have access to Towers Watson Emblem, gave background on GLMs, and carried out exploratory data analysis
- In today's session we will fit a Generalized Linear Model (GLM) on the same LTC Incidence Data, using Towers Watson Emblem

Agenda of today's workshop

- Introduction/Recap
- Partitioning data between modeling/validation, selecting model structure and implications *(15 mins)*
- Fitting a starting model and interpreting results *(30 mins)*
- Assessing additional factors for inclusion *(45 mins)*
- Investigating interactions *(30 mins)*
- Simplifying factors using groups and curves *(1 hour)*
- Validating assumptions and modeling decisions *(30 mins)*
- Testing/comparing predictiveness of model/s *(30 mins)*
- Conclusion and Q&A *(<1 hour)*

Introduction/Recap

There are two goals in fitting a predictive model

- Prediction: to be able to predict responses for future input variables
- Information: to understand the process that associates response variables with input variables

Partitioning data between modeling/validation sets

- We fit the model on one subset of the data (modeling/training data)
- We keep another subset of the data (testing/validation/hold-out data) in reserve to
  - test predictiveness of the model
  - compare the predictiveness of different models
- More on this later
- In this example, we will model on a random three quarters of the data, and use the remaining quarter for validation

Statistical Background to GLMs

GLMs (Generalized Linear Models) are characterized as follows:

Response Variable = Systematic Component + Random Component

y = h(Combination of Factors) + Error

$g=h^{-1}$ is called the LINK function and is chosen to reflect the signal most accurately

Combination of factors is the model structure

Error should reflect underlying process and comes from the exponential family

Statistical Background to GLMs

- Generally accepted standards are good starting points for link functions and error structures

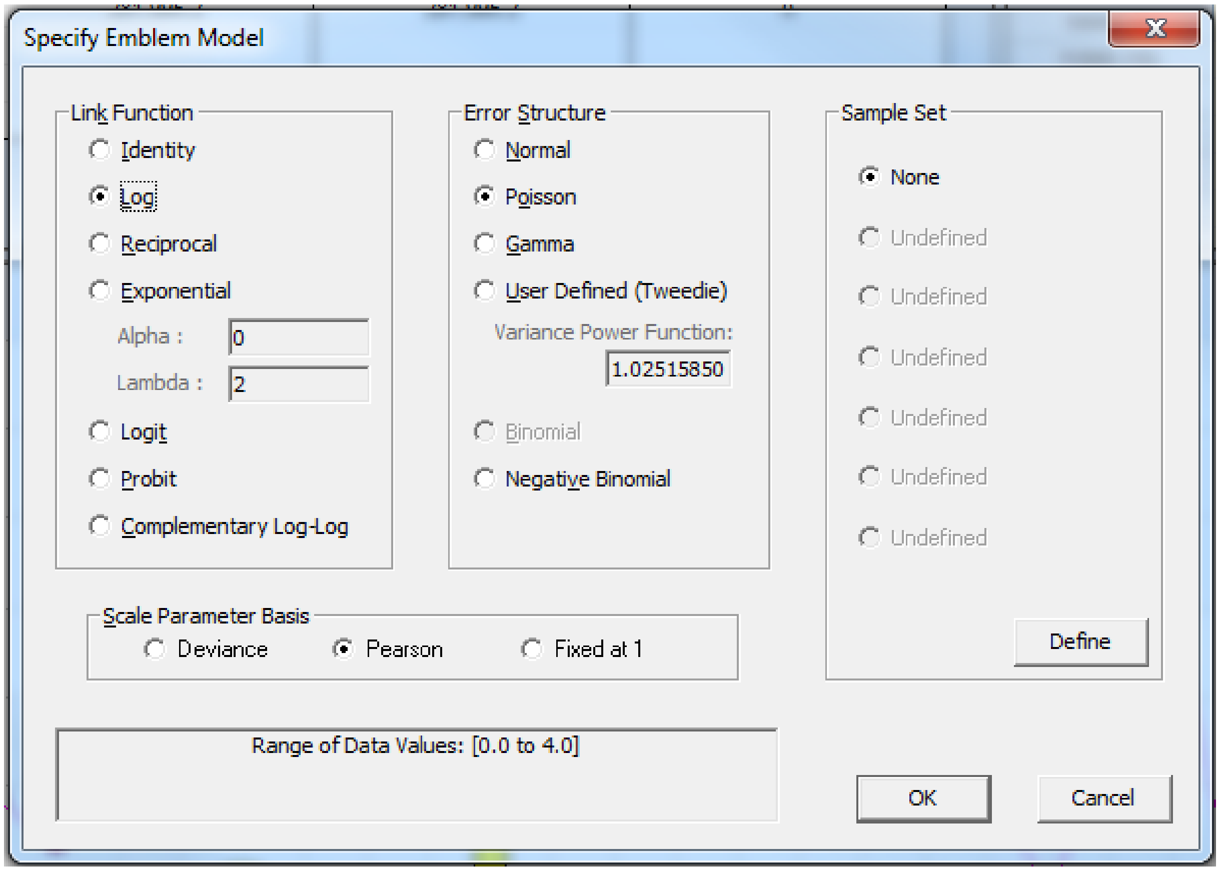
Observed Response	Most Appropriate Link Function	Most Appropriate Error Structure
--	--	Normal
Frequency/Mortality/Incidence	Log	Poisson
Severity/Utilization	Log	Gamma
Pure Premium	Log	Tweedie
Retention/Conversion/Termination/ Cross-sell/Response Rate	Logit	Binomial

Selecting model structure and implications

- We will be using Log link function and Poisson error structure
  - Conceptually this means that we are modeling events per unit time
  - This is a standard choice for modeling claims frequency in P&C insurance and has been used for LTC incidence
- A consequence of this is that the model is multiplicative

Selecting model structure and implications

- We define training and testing data and model link and error structure on the "Specify Emblem Model" window:

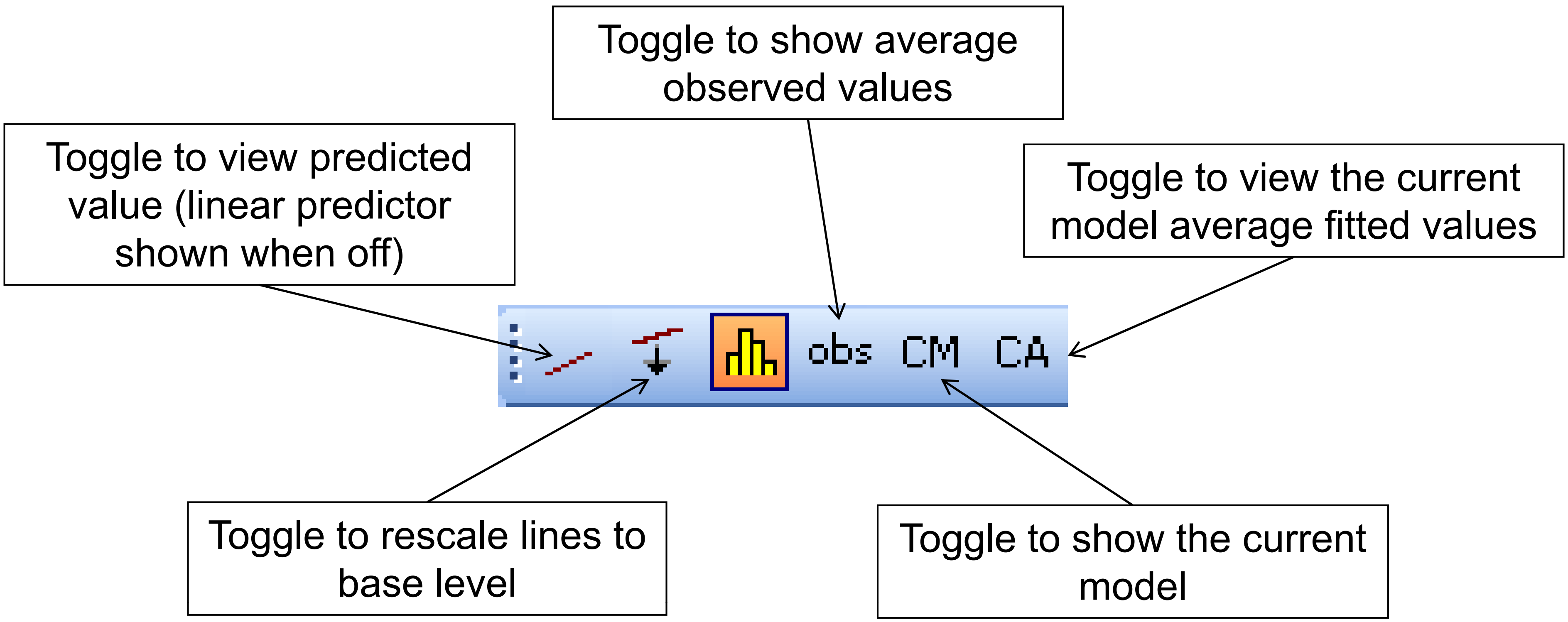


Fitting a starting model

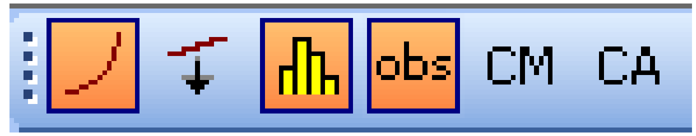
- There are various ways to choose a starting model
- These include
  - Include variables from an existing assumption
  - Include variables you suspect will be predictive
  - Stepwise Regression
  - Another analysis, such as CART, to rank “importance” of factors
- We will start from a one variable model (DurYear)
  - This is not standard, but allows us to explain a few basic concepts

Basic Concepts

- The Graph toolbar contains buttons which allow the user to customize the graph for the selected variable



Basic Concepts: Observed (Actual)

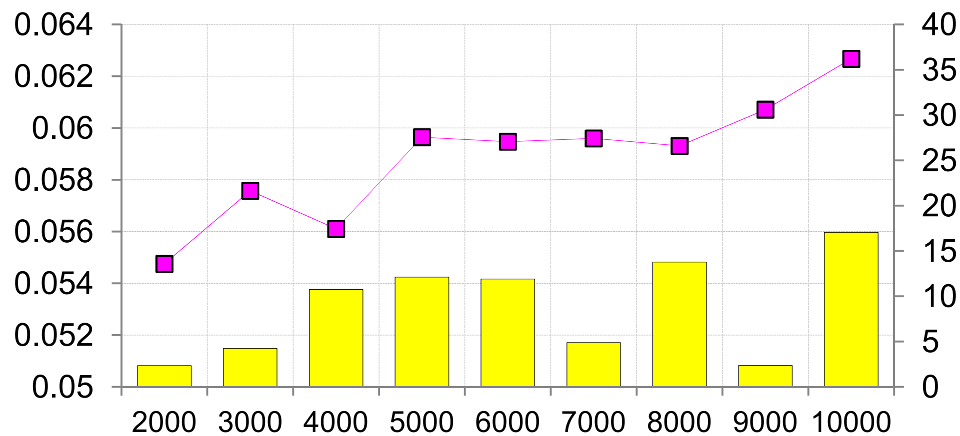


- The **obs** button toggles the average observed values, which are weighted by exposure.
- Within each variable level,

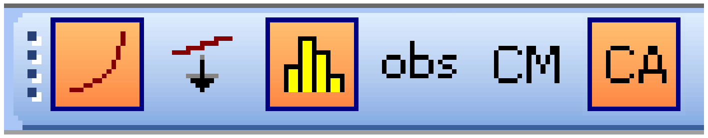
$$\text{Average Observed Value} = \frac{\text{Weighted Response}}{\text{Total Weight}}$$

Annual Mileage	Total Weight	Weighted Response	Average Observed Value
2000	23,269	1,274	0.0548
3000	42,153	2,427	0.0576
4000	106,574	5,979	0.0561
5000	119,996	7,157	0.0596
6000	117,888	7,011	0.0595
7000	48,426	2,886	0.0596
8000	136,342	8,086	0.0593
9000	23,374	1,419	0.0607
10000	168,833	10,581	0.0627

Annual Mileage – Observed Average



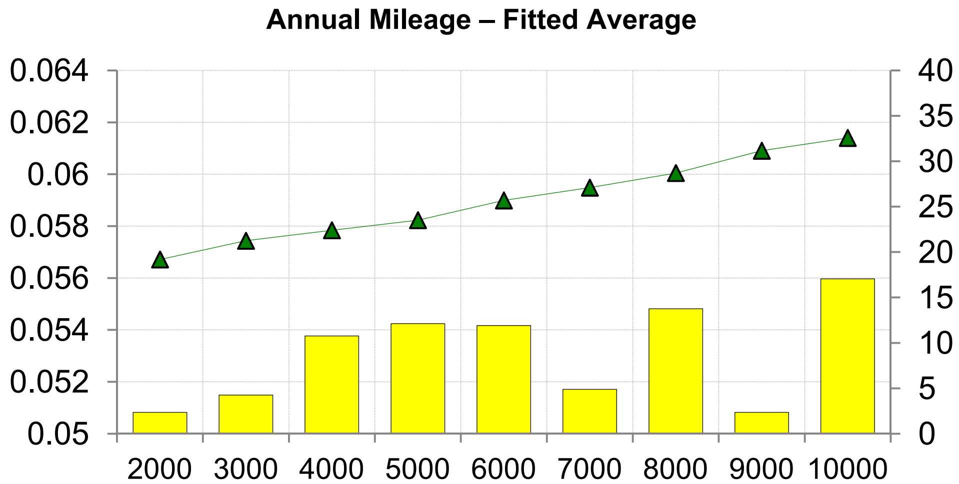
Basic Concepts: Fitted (Expected)



- The **CA** button toggles the average fitted values of the current model.
- Within each variable level,

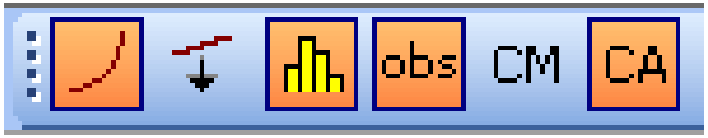
$$Average\ Fitted\ Value = \frac{Weighted\ Fitted\ Value\ of\ Current\ Model}{Total\ Weight}$$

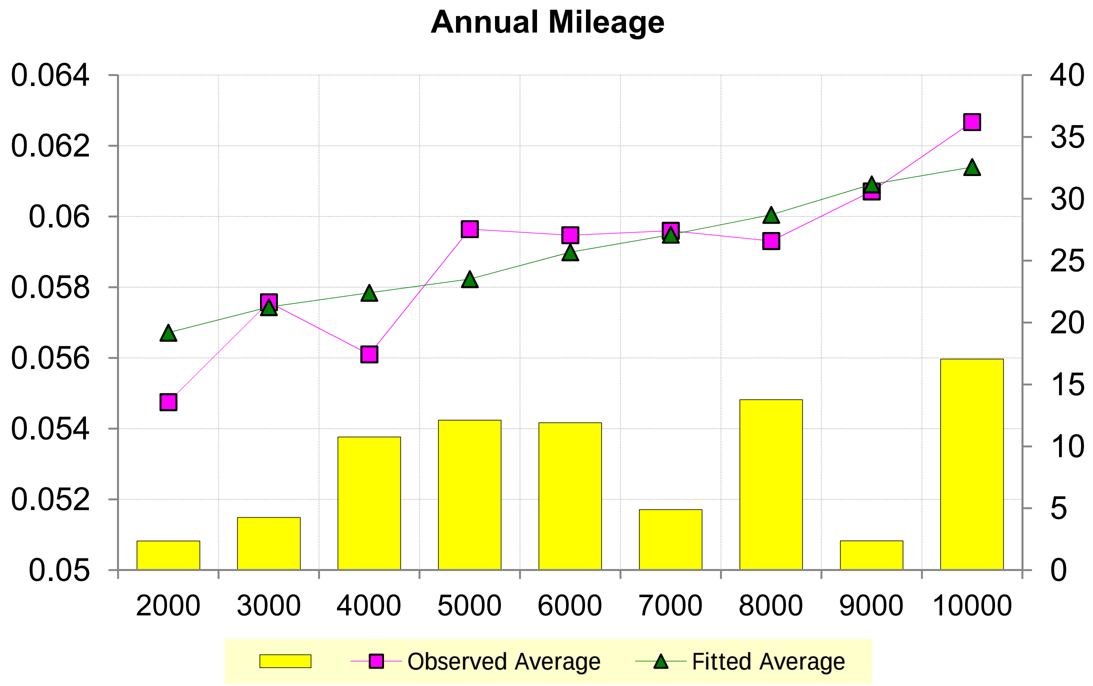
Annual Mileage	Total Weight	Weighted Fitted Value	Average Fitted Value
2000	23,269	1320	0.0567
3000	42,153	2421	0.0574
4000	106,574	6165	0.0578
5000	119,996	6987	0.0582
6000	117,888	6954	0.059
7000	48,426	2881	0.0595
8000	136,342	8187	0.06
9000	23,374	1424	0.0609
10000	168,833	10366	0.0614



Basic Concepts: Actual vs Expected

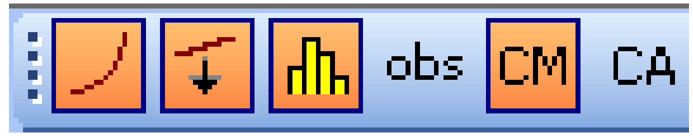
- The **obs** and **CA** buttons are often toggled together in order to examine whether the current model is in balance with the data.

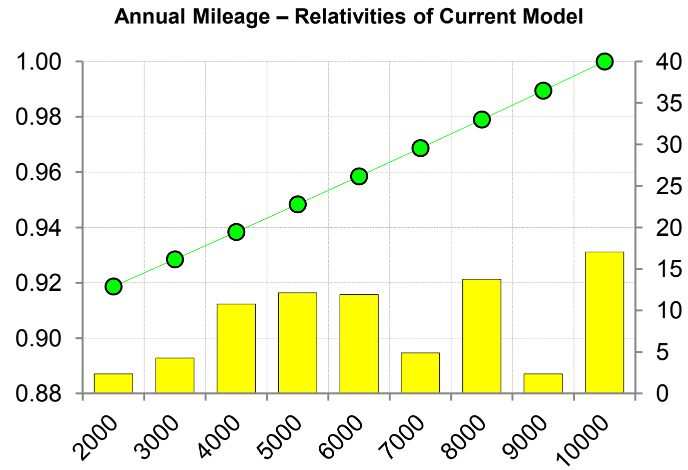




Basic Concepts: Relativities

- The **CM** button with **predicted value** and **rescale** buttons toggle the relativities of the current model.
- The rescale button standardizes each predicted value by dividing by the predicted value of the model's base parameter.
- In the following example, the base level is 10000 for Annual Mileage:

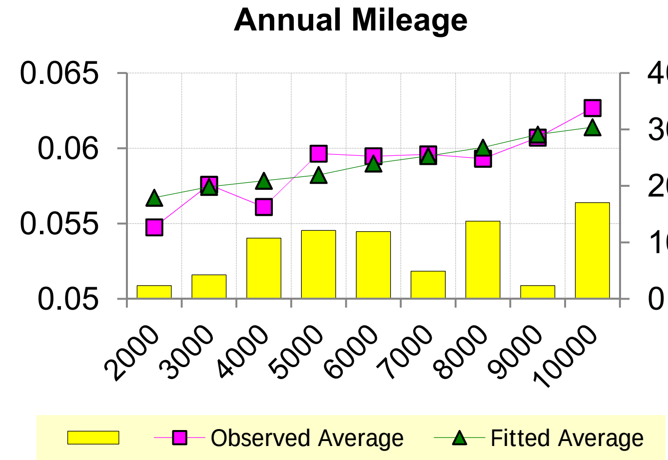


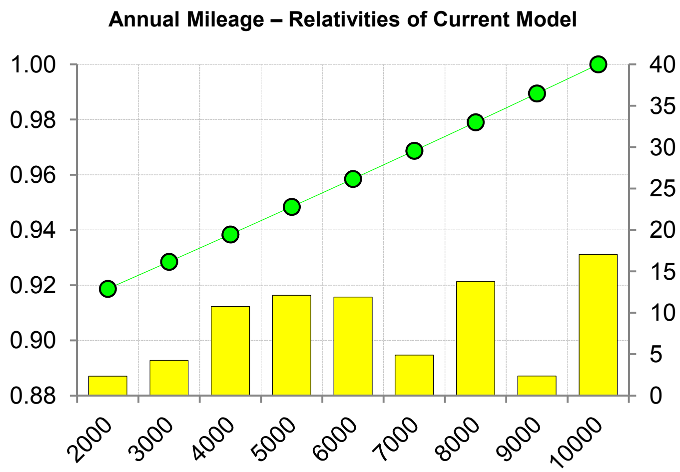


Annual Mileage	Predicted Values of Current Model	Relativities of Current Model
2000	0.0496	0.9187
3000	0.0501	0.9285
4000	0.0506	0.9384
5000	0.0512	0.9484
6000	0.0517	0.9585
7000	0.0523	0.9687
8000	0.0528	0.9790
9000	0.0534	0.9895
10000	0.054	1.0000

Basic Concepts: Summary

- Crimson Line (Obs): actual or observed effect
- Dark Brown Line (CA): fitted or expected effect



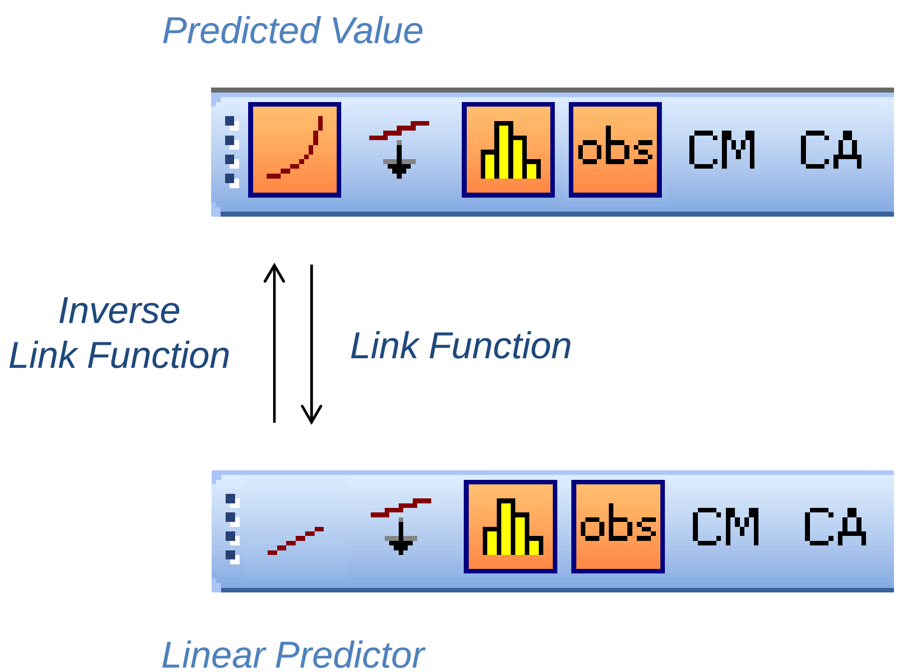


- Green Line (CM): effect of this variable in this model (standardizing for all the other variables in the model)

- Difference between observed and modeled effects is owing to impact of other factors in the model

Basic Concepts: Predicted Values vs Linear Predictor

- The **predicted value/linear predictor** button toggles whether the plotted lines use the inverse link function (*for predicted value*) or the link function (*for linear predictor*).
- For a Log link function,
  - Viewing the predicted values of the current model's parameter values means exponentiating them



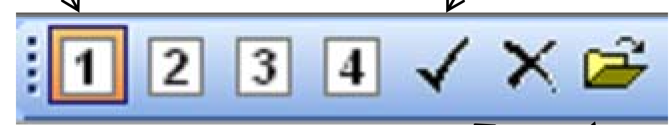
Linear Predictor

y = h(Combination of Factors) + Error

Fitted Values

Basic Concepts: Reference Model

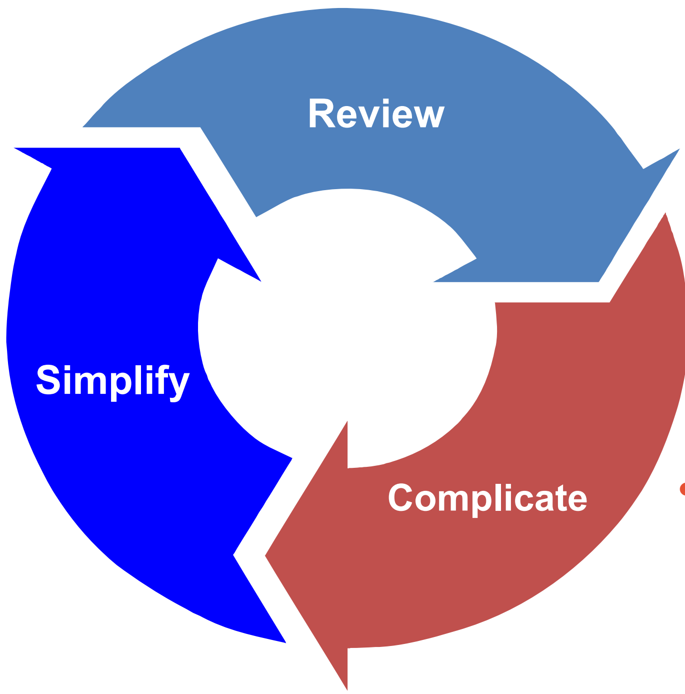
- Defined using the Define Reference Model icon
- Possible to define up to 4 reference models
- Can be compared against the current model using the statistics tab
- Trend lines corresponding to the reference model can be added to the main graph
- Can be reloaded using the Reload Reference Model icon



- Set your current model as reference model 1

Modeling Process

- Building the model is an iterative process



- Review modeling decisions

- Exclude effects
- Simplify effects with groups and curves

- Include new effects

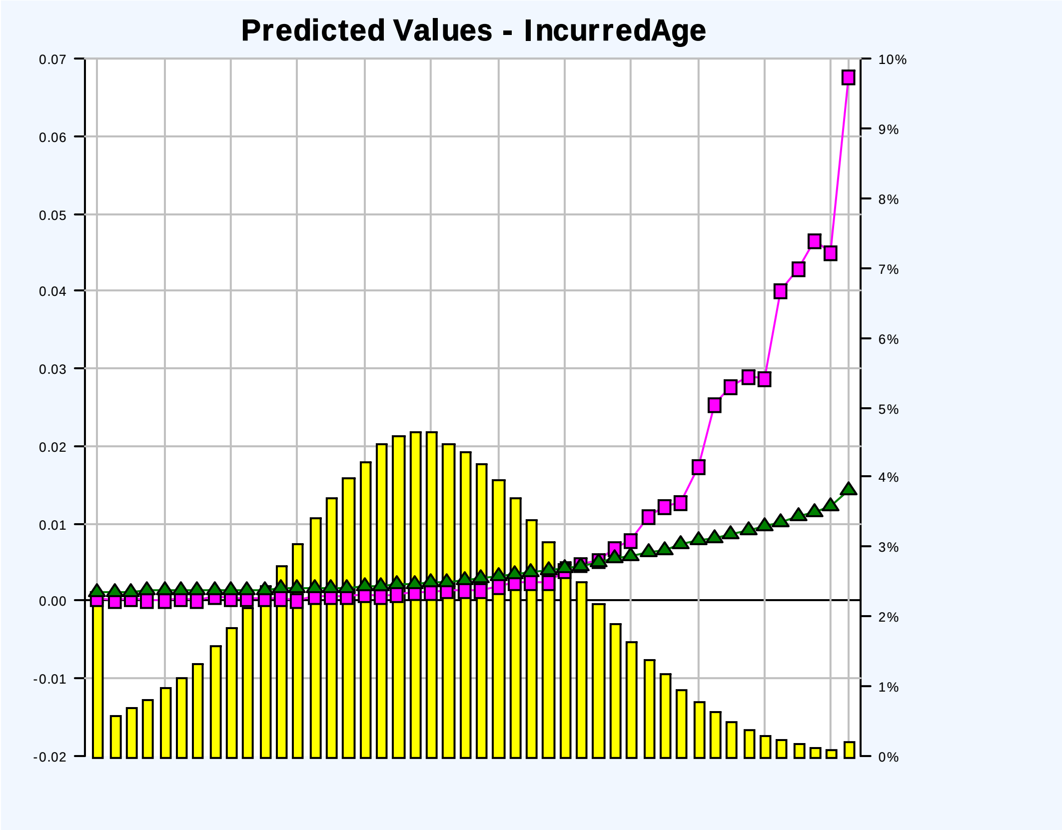
Testing for Factor Inclusion/Exclusion

- We use the following tests to decide which variables to include in our model
  - Balance Test (comparing Actual vs Expected)
  - Confidence Intervals of Parameter Estimates
  - Statistics (Chi-square, AIC, BIC)
  - Consistency of Patterns
  - Sense Check/Judgment

Testing for Factor Inclusion/Exclusion

Balance Test

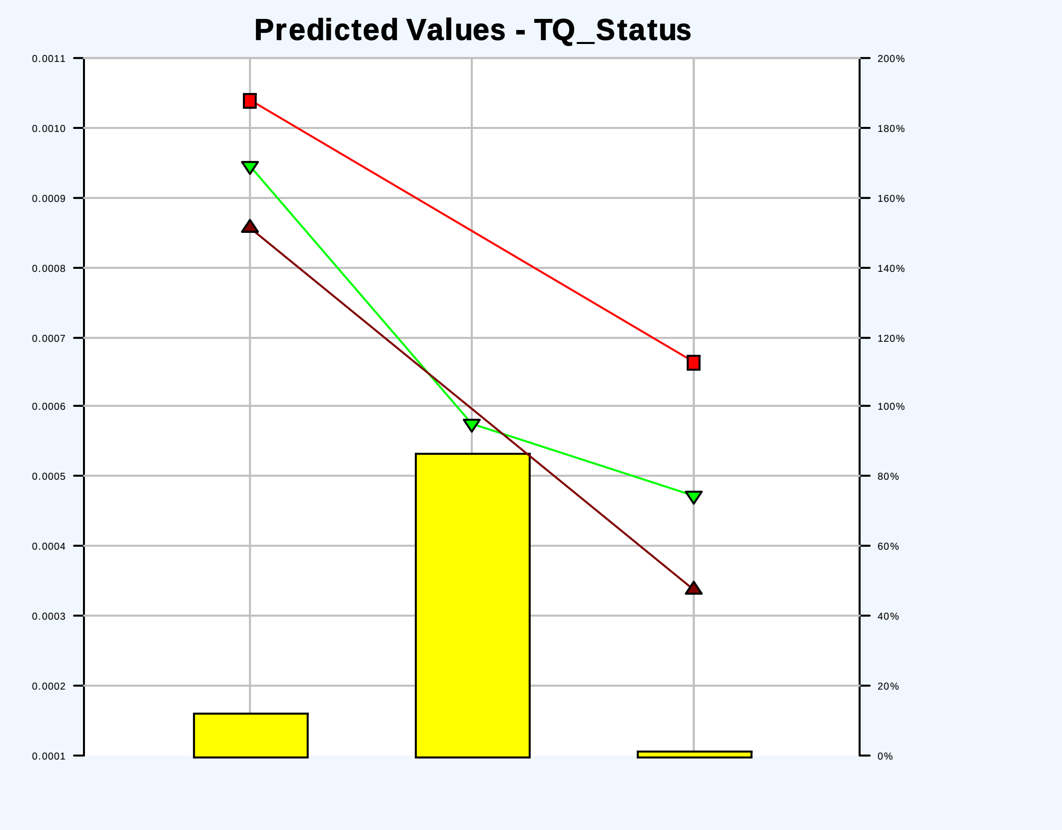
- If Actual and Expected are similar on a univariate basis, we say that the model is "in balance" by this variable
- If a variable is out of balance, we should investigate adding the variable



Testing for Factor Inclusion/Exclusion

Confidence Intervals of Parameter Estimates

- If the confidence intervals of all levels of a variable include the base, the variable could be considered for exclusion from the model



Testing for Factor Inclusion/Exclusion

Statistics

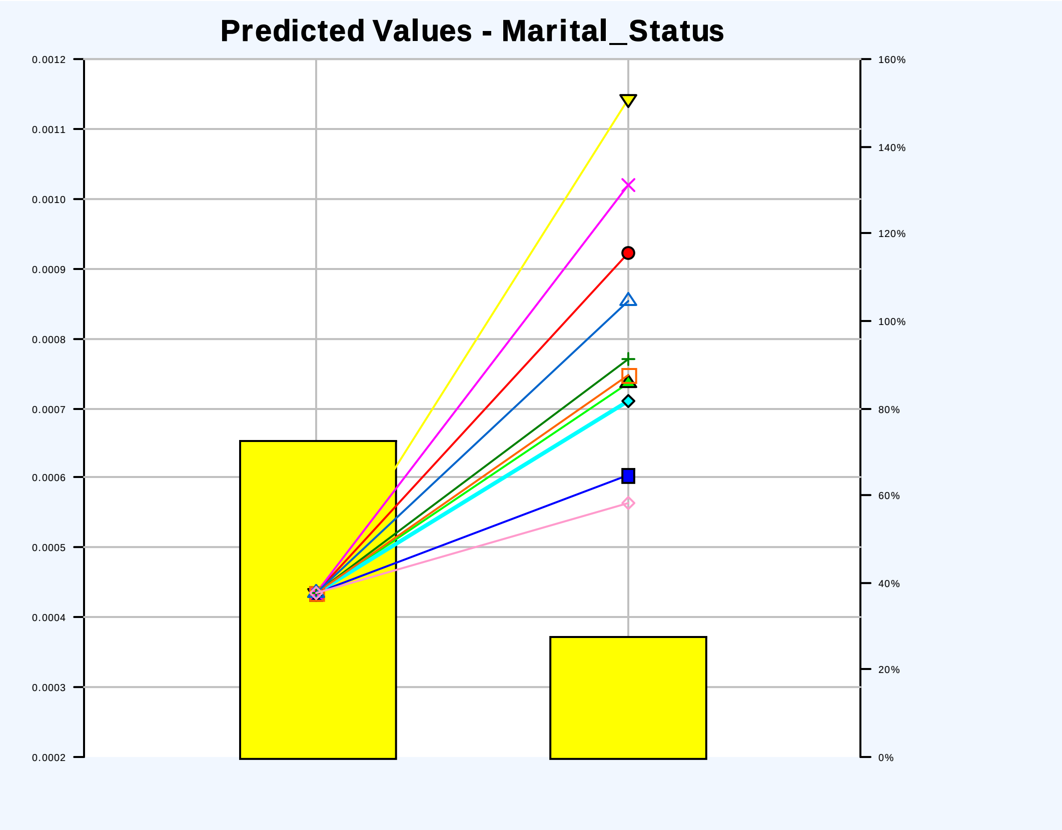
- We can compare statistics between current and reference models
- Chi-square:
  - Allows a hypothesis test of nested models
  - The closer to zero, the stronger the result of the test
  - 5% is a common cut-off
- Information criteria (AIC, BIC):
  - Allow a comparison of two models (not necessarily nested)
  - These are a trade-off between fit to the data and complexity of the model
  - The lower the criteria, the better
  - BIC punishes inclusion of additional parameters more than AIC

	Current Model	Reference Model	Difference
Model Label	(none)*	(none)*	
Sampling	Training	Training	
Model Description	Mean + DurYear + IncurredAge	Mean + DurYear	+ IncurredAge
Zero Weighted	988,690	988,690	0
Fixed or Simple Alias	0	0	0
Complex Alias	0	0	0
Fitted Parameters	63	18	45
Deviance	31,185.93	34,857.72	-3,671.794
Chi Squared Percentage		Sub-Model	0.0%
AIC	34,074.63	37,656.42	-3,581.794
BIC	34,887.48	37,888.67	-3,001.182
Fitting Result	Converged OK	Converged OK	

Testing for Factor Inclusion/Exclusion

Consistency of Patterns

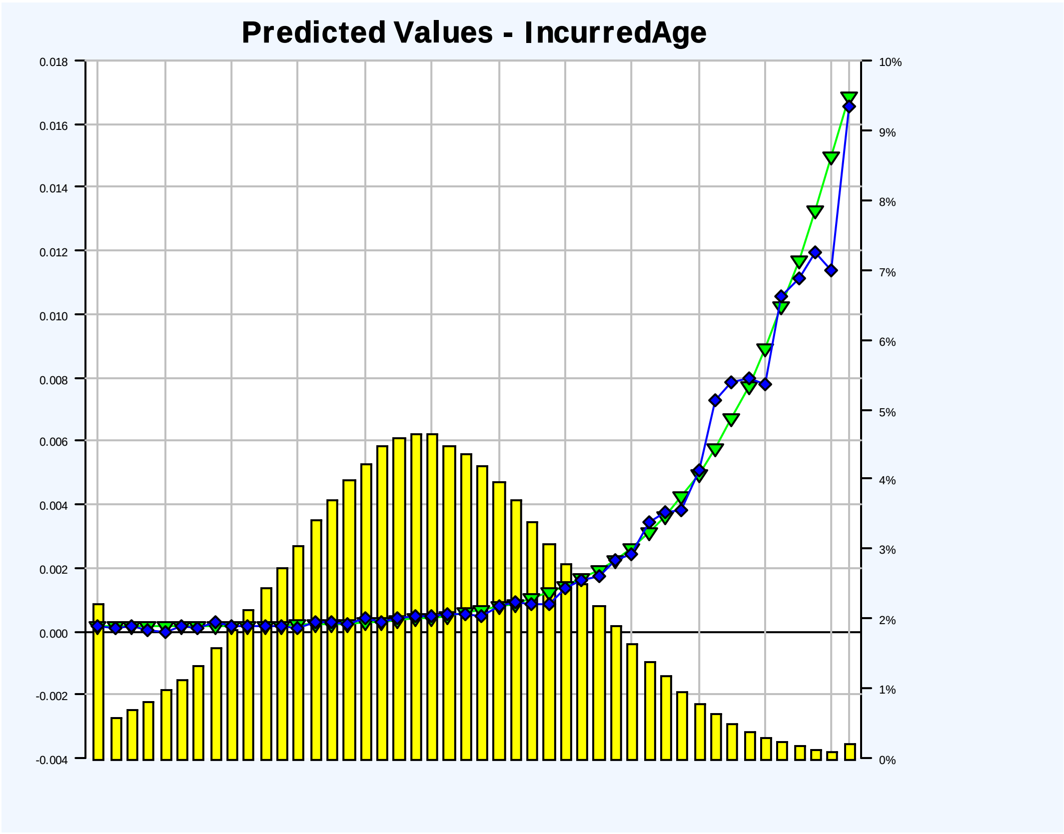
- If a variable has parameters which are consistent across a random split of the data or some other factor (such as time) it gives us more confidence that the parameters are not being driven by some isolated part of the modeling data



Testing for Factor Inclusion/Exclusion

Sense Check/Use of Judgment

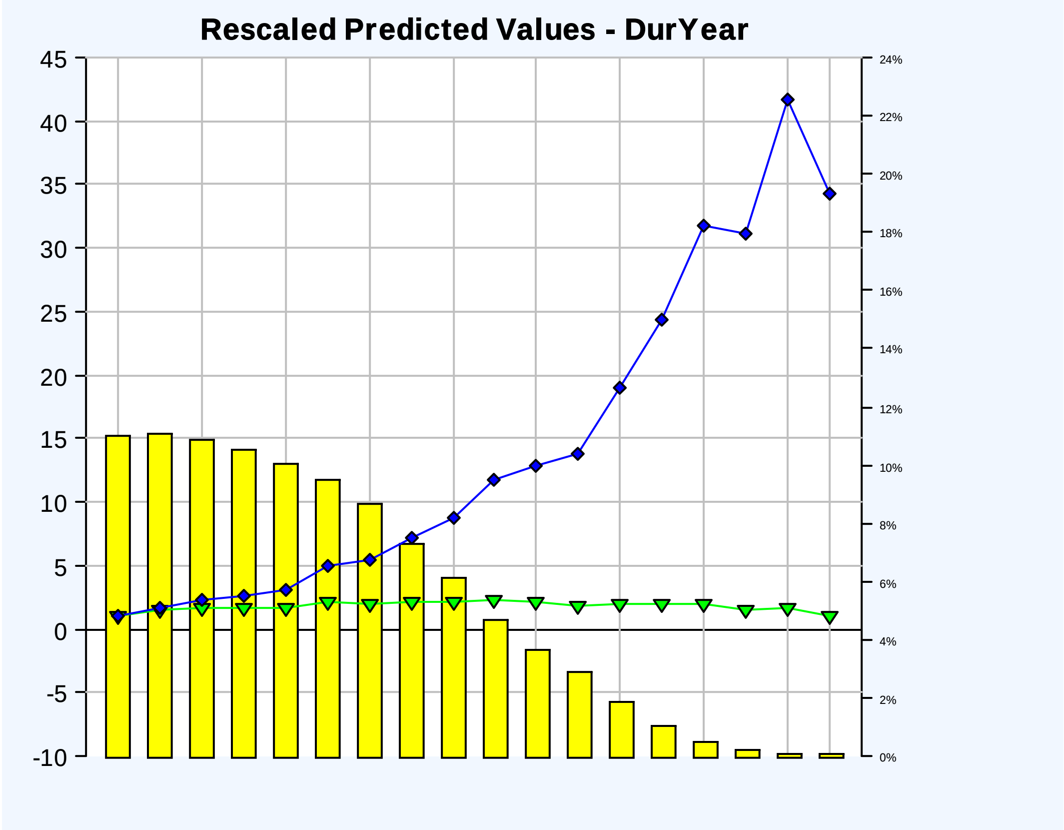
- It is preferable to be able to explain the effects included in the model
- Ask yourself: does the effect make sense?



- We will try adding Incurred Age to the model

Testing for Factor Inclusion/Exclusion

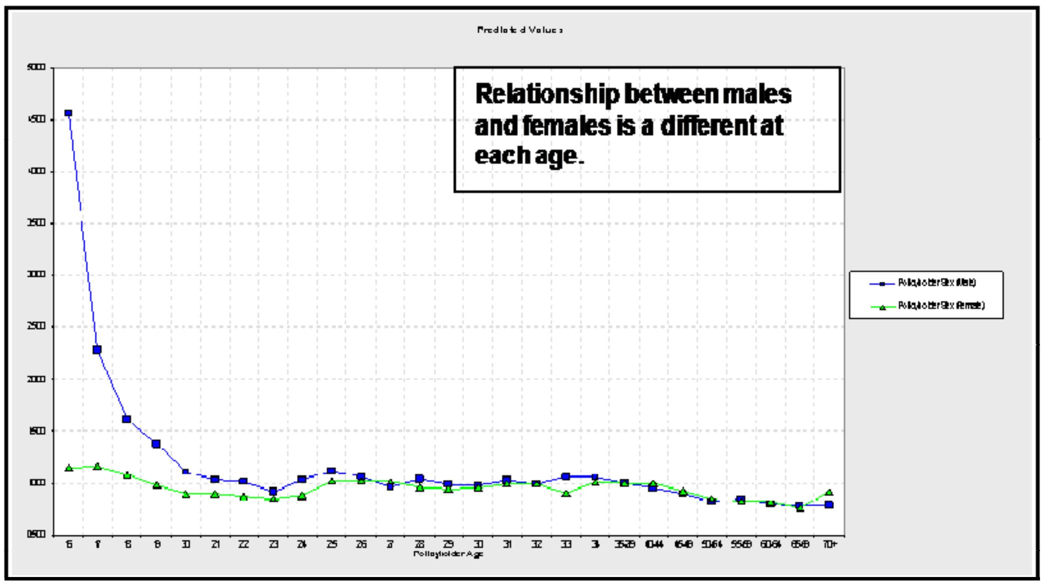
- We can visualize the impact on the Duration effect of adding Incurred Age
- The green line shows the effect of duration once incurred age is taken into account
- Remember that one of the goals of predictive modeling is to understand the process



- Try adding other factors to your model

Investigating interactions

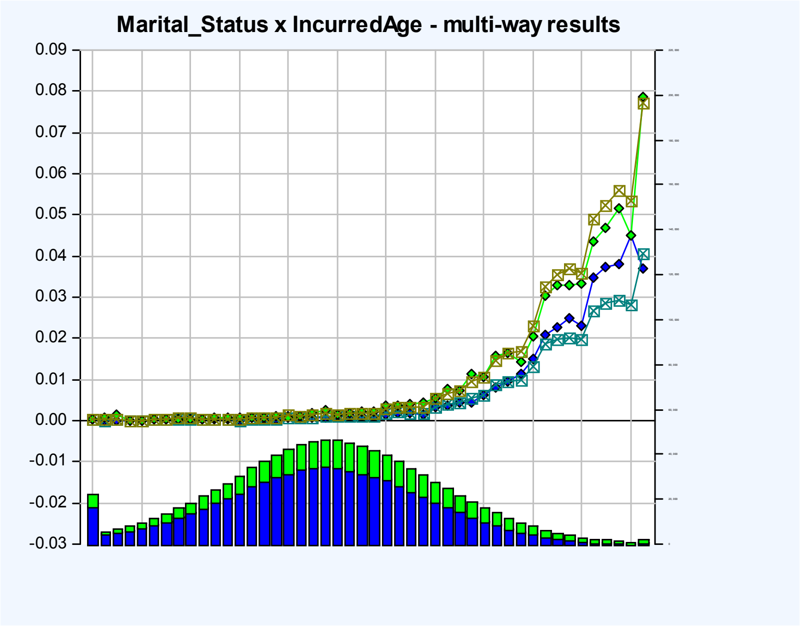
- Sometimes it is not enough to include two variables in the model- it is necessary to include their combination
- This is when the impact of one variable depends on the level of another variable
- This is called an <u>interaction</u>
- In auto insurance, the canonical example is age x gender

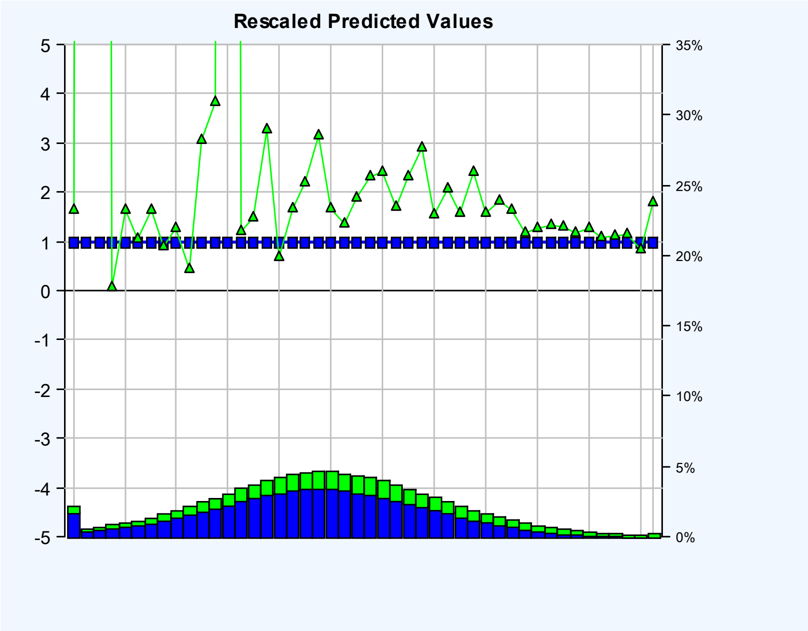


- Interactions can be found by
  - Inspection: looking to see where combinations of factors are “out of balance”
  - Calculation: calculating combinations of factors that are out of balance
  - Distortion: of an existing model

Investigating interactions

- Here we will investigate Incurred Age and Marital Status
- This pair of factors is out of balance
- If we add the interaction into the model, we can see how the Incurred Age effect varies for each level of Marital Status (i.e. age has a different impact for singles and marrieds)





Simplifying factors using groups and curves

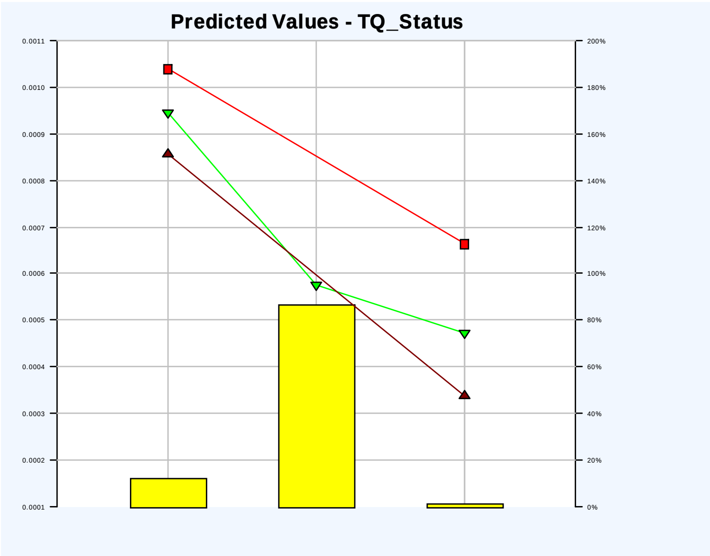
Why do we simplify models?

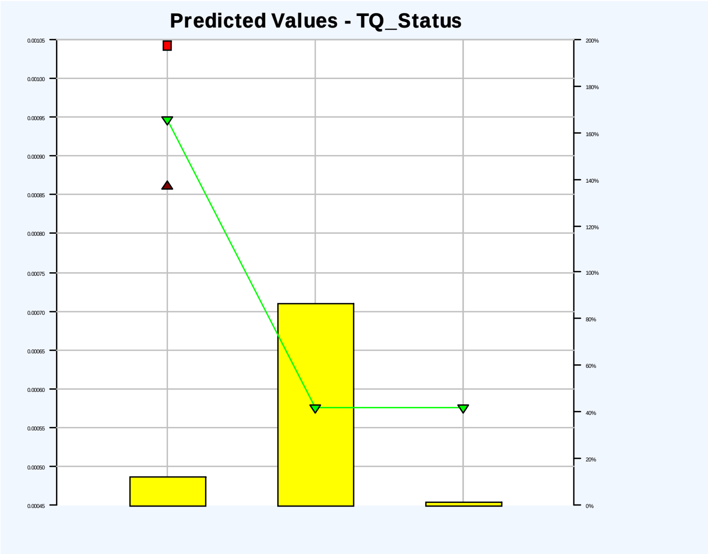
- For statistical reasons:
  - A parsimonious model is better
  - Einstein: “A model should be as simple as possible but no simpler”
- For business/conceptual reasons:
  - Ordinal variables (age, duration, coverage amounts) should in most cases vary smoothly
- We will carry out two kinds of simplifications:
  - Groupings for categorical factors
  - Curves for ordinal factors

Simplifying factors using groups and curves

Grouping categorical variables

- If two levels have similar parameters, it may make sense to group them
- It may also make sense to group "small" levels with the base, or some other reasonable level

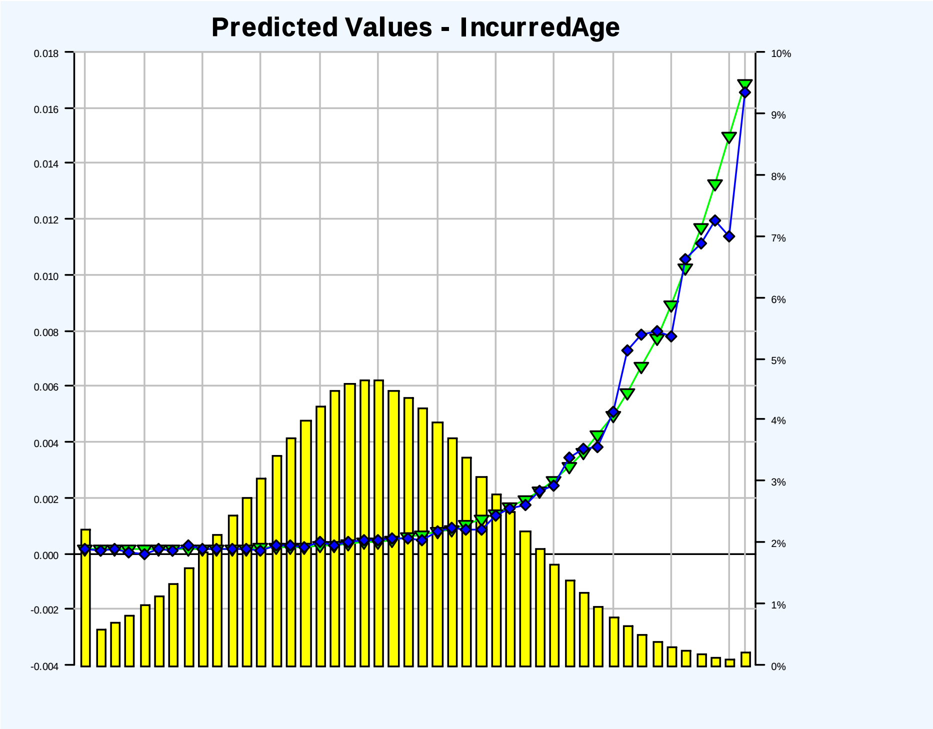




Simplifying factors using groups and curves

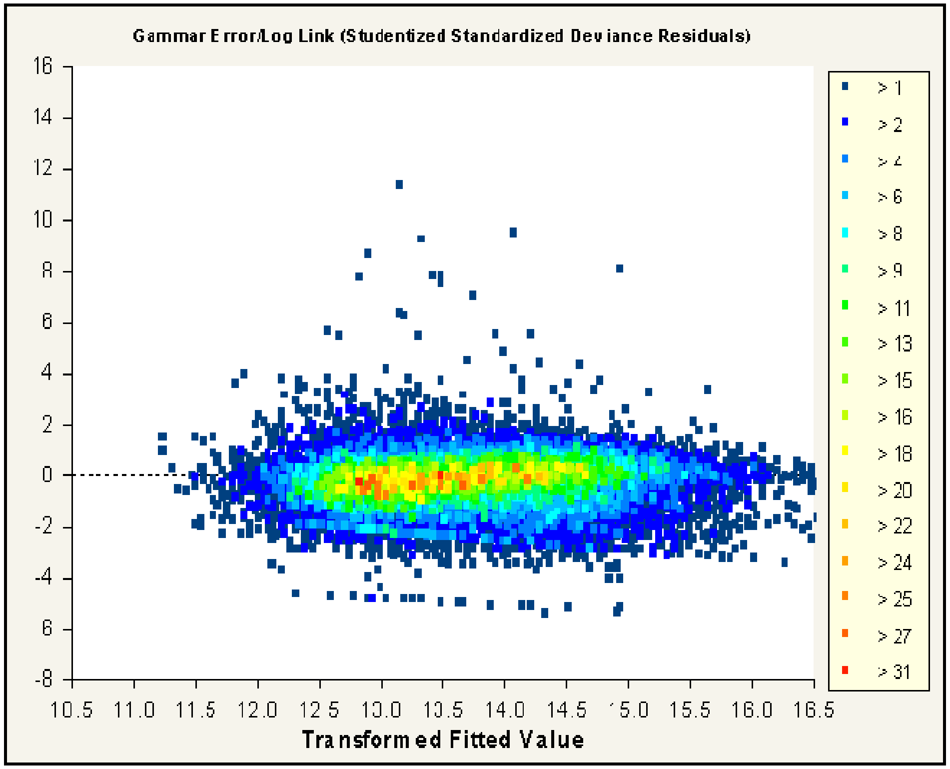
Ordinal variables can be simplified with curves

- Rather than having one parameter per level, it may make sense to include a function of variable levels in the model
- Common examples are
  - Polynomials: a*x, a*x+b*x^2 etc.
  - Logs: ln(x)
- This makes the model more parsimonious, and means that effects are smooth, which can make more sense intuitively, and be more useful for business applications



Validating assumptions and modeling decisions

- Residual Plots
  - We should run a residual plot in order to check that our model assumptions are appropriate
  - Plot should be symmetrical about the vertical axis, with no obvious pattern
  - Depending on the data being modeled, this is more or less complicated



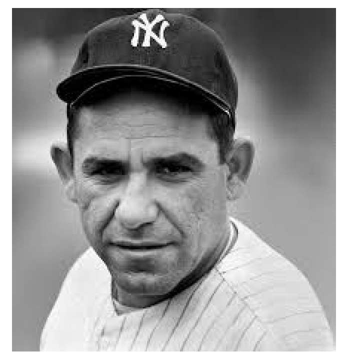
- Revisiting Modeling Decisions
  - We should review all of our modeling decisions, including factor inclusion/exclusion, simplifications and interactions
  - We will do this using Model Manager, after saving our model

Testing/comparing predictiveness of model/s

- Our two goals in predictive modeling were
  - Prediction: to be able to predict responses for future input variables x
  - Information: to understand the process that associates response variables with input variables

"Prediction is difficult, especially when you want to predict the future"

Attributed to Niels Bohr…

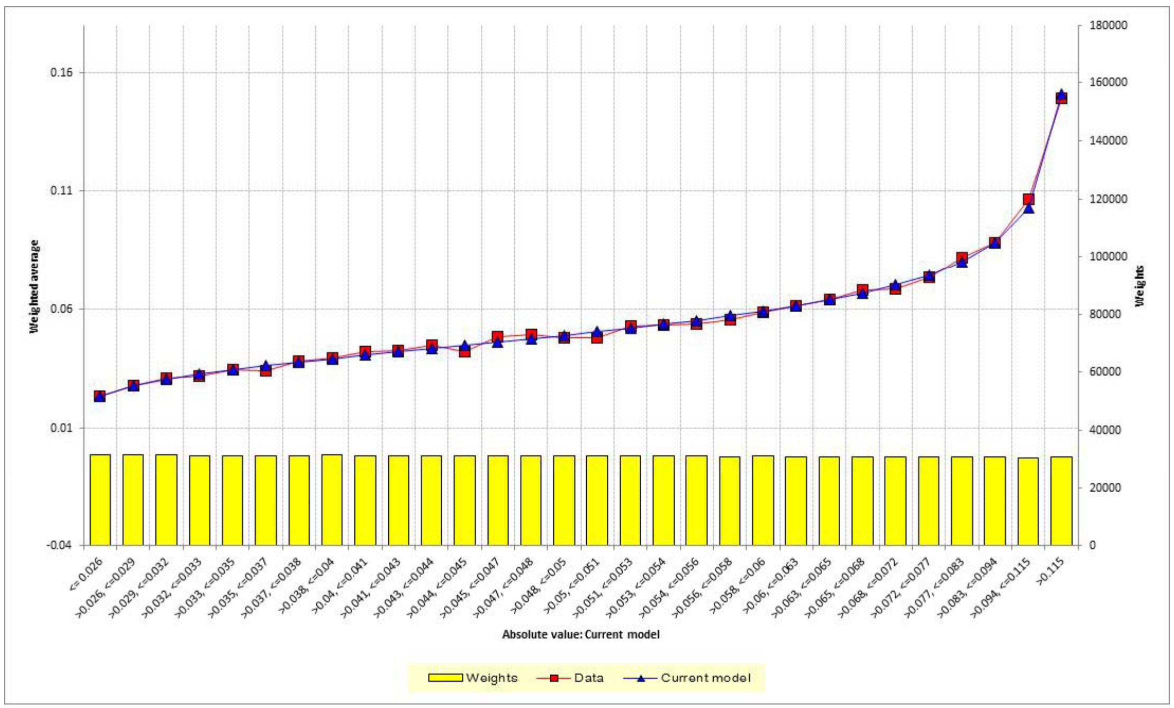


…and Yogi Berra

- We now show:
  - How to test predictiveness of a model on hold-out data
  - How to compare predictiveness of models on hold-out data

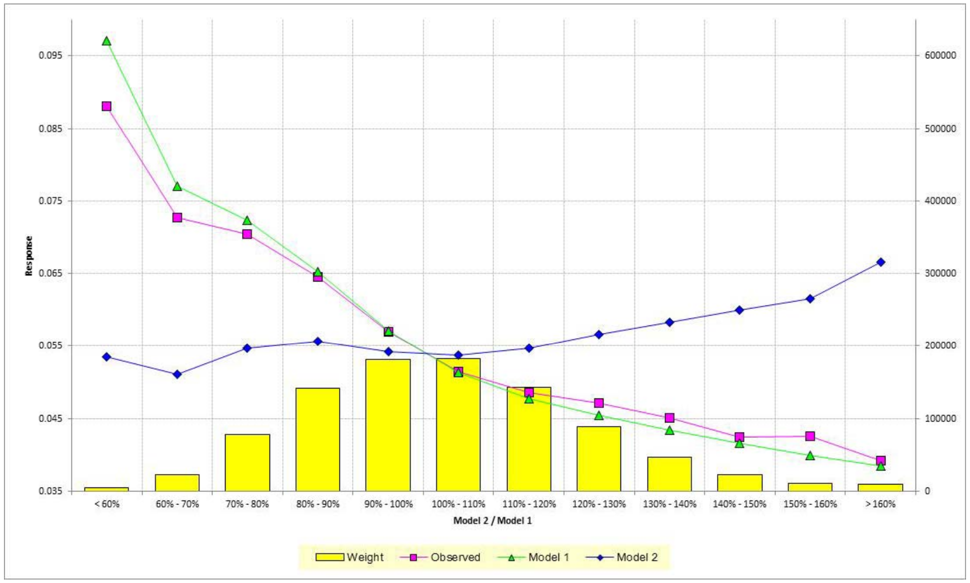
Testing/comparing predictiveness of model/s

- We can compare Actuals to Expected on hold out data
  - On a univariate basis, by different variables. We expect the lines to be close for well-populated levels
  - In a Lift Chart, where the horizontal axis is the model fitted value.



Testing/comparing predictiveness of model/s

- If we want to compare two models, we can
  - Calculate the fitted value for both on the hold-out data,
  - Calculate the ratio of fitted values
  - Use this as the horizontal axis of a chart
  - In each interval of “model difference” we can calculate the Actual and the Expected according to each model
- This tells us
  - How different the models are
  - Where they are different, which makes better predictions



Conclusion

- Once we are happy with our model, we can use it to make predictions
- Output will look something like this

Base	0.0004								
Gender		**Incurred Age**		**Duration**		**Marital Status**		**Premium Class**	
Female	1.0000	35-	0.1171	1	1.0000	Married	1.0000	Preferred	0.8682
Male	0.7560	36	0.1326	2	1.4224	Single	2.1227	Standard	1.0000
		37	0.1484	3	1.8732			Substandard	1.1006
		38	0.1641	4	2.3096				
		39	0.1796	5	2.6933				
		40	0.1947	6	2.9984				
		41	0.2092	7	3.2136				

Base	Gender	Incurred Age	Duration	Marital Status	Premium Class	Expected Mortality (per '000)
	Male	51	5	Single	Preferred	
0.0004	0.7560	0.3293	2.6933	2.1227	0.8682	0.4775

50	0.3171	16	2.8584
51	0.3293	17	2.7583
52	0.3425	18	2.6561

Conclusion

- Today we fit a GLM of LTC Incidence on real data
- This allowed us to
  - Understand incidence (what variables, and combinations of variables, impact on incidence- and the effect of each)
  - Make predictions about the incidence to be expected for certain combinations of variables
- If you have any questions, please contact us:
  - Benjamin.Williams@willistowerswatson.com
  - Matt.Morton@LTCG.com
- A useful reference for GLMs is:
  - A Practitioner's Guide to Generalized Linear Models, by Anderson, Feldblum, Modlin, Schirmacher, Schirmacher and Tandi
- We thank you for your participation in the session, hope that you have found it interesting, and wish you a safe journey home